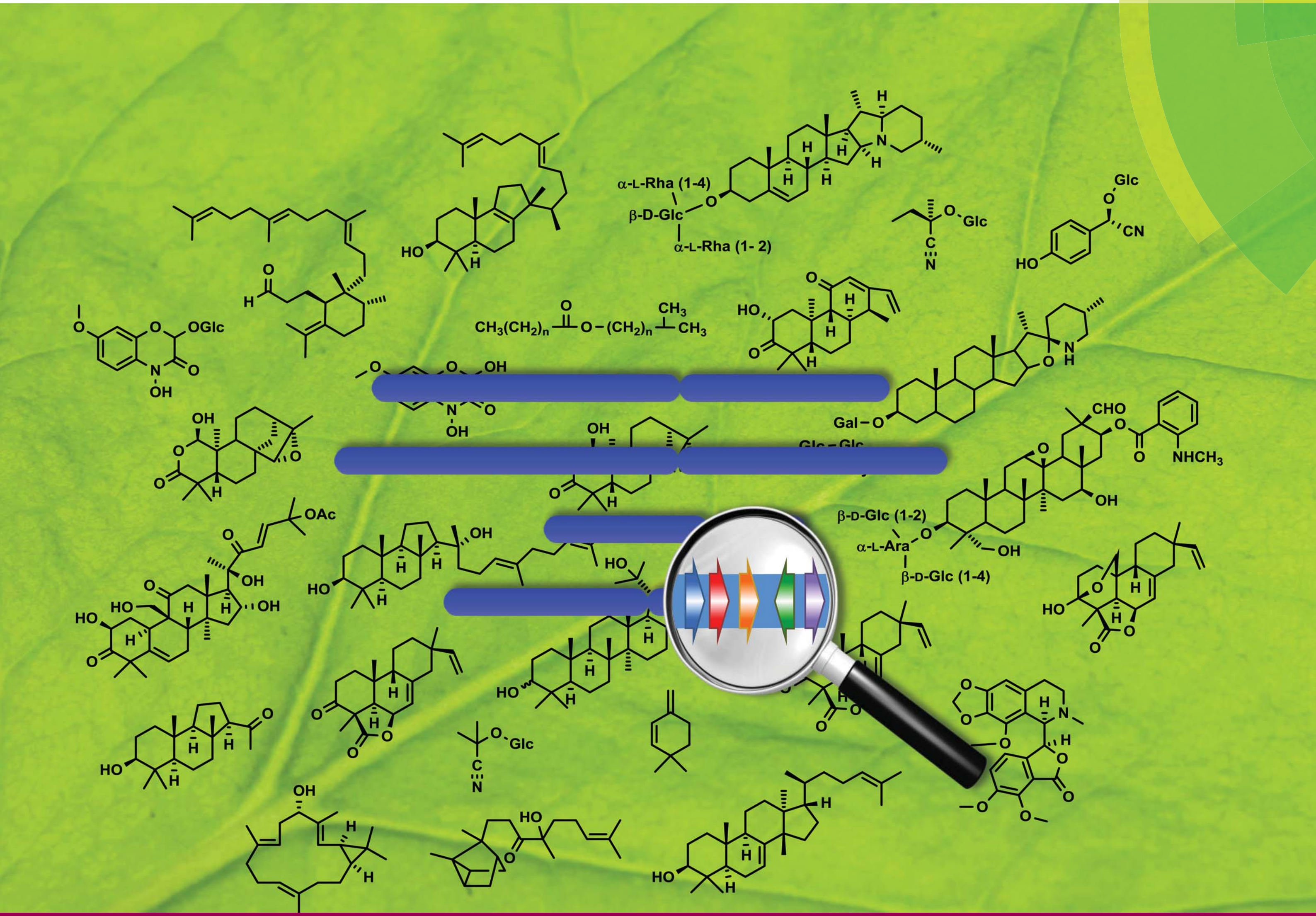
Volume 33 | Number 8 | August 2016 | Pages 907–1022

Natural Product Reports

www.rsc.org/npr



Themed issue: Synthetic biology and bioinformatics

ISSN 0265-0568

ROYAL SOCIETY OF CHEMISTRY

REVIEW ARTICLE
Marnix H. Medema and Anne Osbourn
Computational genomic identification and functional reconstitution of plant natural product biosynthetic pathways

175 YEARS

Natural Product Reports

REVIEW

View Article Online
View Journal | View Issue

Cite this: *Nat. Prod. Rep.*, 2016, **33**, 951

Computational genomic identification and functional reconstitution of plant natural product biosynthetic pathways

Marnix H. Medema*[a] and Anne Osbourn*[b]

Covering: 2003 to 2016

The last decade has seen the first major discoveries regarding the genomic basis of plant natural product biosynthetic pathways. Four key computationally driven strategies have been developed to identify such pathways, which make use of physical clustering, co-expression, evolutionary co-occurrence and epigenomic co-regulation of the genes involved in producing a plant natural product. Here, we discuss how these approaches can be used for the discovery of plant biosynthetic pathways encoded by both chromosomally clustered and non-clustered genes. Additionally, we will discuss opportunities to prioritize plant gene clusters for experimental characterization, and end with a forward-looking perspective on how synthetic biology technologies will allow effective functional reconstitution of candidate pathways using a variety of genetic systems.

Received 15th March 2016

DOI: 10.1039/c6np00035e

www.rsc.org/npr

1. Introduction

Plants produce a huge array of natural products. Humans have relied on plants as sources of traditional medicines and drugs, dyes, colours, flavours, fragrances and agrochemicals throughout history. The compounds that have been exploited so far, however, represent only the tip of the iceberg in terms of the chemical diversity that is represented within the plant kingdom. Plant natural products are often lineage-specific, having evolved within narrow taxonomic groupings. They are normally produced only at certain growth stages in particular tissues, or in response to environmental stimuli (for example, challenge with pests, pathogens or elicitors), and are often found only at low concentrations within complex mixtures in plant extracts. These factors frequently cause problems in obtaining and/or cultivating the material and extracting and purifying the compounds of interest in useful quantities, for either research purposes or for commercial production. For these reasons, it difficult to access natural products from plants.

Identification of the genes involved in the biosynthetic pathways for the production of these molecules is a promising remedy, as it potentially allows heterologous expression of a pathway to acquire higher yields. Over seventy genome sequences have now been determined for different plant species and, in addition, a wealth of transcriptome data is available. Based on superficial analysis of these genomic resources, it is clear that plant genomes encode the capacity for an enormous amount of metabolic complexity. However, interpretation of this information and translation of predicted sequences into enzymes, pathways and products represents a major challenge. Systematic analysis of the metabolic capacity of particular plant species using transcriptomics and metabolomics has proved to be a very useful strategy for identifying candidate genes implicated in the biosynthesis of different types of natural products.

[a] *Bioinformatics Group, Wageningen University, Wageningen, The Netherlands. E-mail: marnix.medema@wur.nl*

[b] *Department of Metabolic Biology, John Innes Centre, Norwich Research Park, Norwich, UK. E-mail: anne.osbourn@jic.ac.uk*

If one or more genes encoding biosynthetic enzymes for the synthesis of a compound of interest have been identified, then this process can be guided by using the genes for characterized enzymes as bait in co-expression analysis, in concert with knowledge of the structure of the compound, to inform on the other enzymes that may be involved in the pathway.[1–3] Alternatively, if candidate genes are identified solely on the basis that they are co-expressed and predicted to encode enzymes of specialized metabolism, but the pathway and the nature of the end product are unknown, then untargeted metabolomics can be a powerful tool to discover entirely new pathways and chemistries.[3]

In microbes, widely used bioinformatics-based approaches to discover new metabolic pathways are based on the identification of physically clustered groups of genes termed 'biosynthetic gene clusters' (BGCs).[4] Intriguingly, it has recently been discovered that in plants the genes for a number of biosynthetic pathways are also encoded in operon-like gene clusters, which may facilitate co-regulation and stable co-inheritance.[5–7] For the most part, these clustered pathways appear to have evolved relatively recently in evolutionary time within narrow taxonomic lineages and are not a consequence of horizontal gene transfer from microbes. This defies the assumption that gene ordering on plant chromosomes is more or less random. It also has an important practical implication: it potentially allows the straightforward identification of biosynthetic pathways from genome sequences, just like in bacteria and fungi. A recent computational study by Chae *et al.*[8] of the genomes of *Arabidopsis thaliana*, rice, soybean and sorghum indicated that genes associated with metabolism are indeed more often clustered than expected by chance in these species, and that the observed clusters of metabolic genes in *A. thaliana* and soybean are significantly enriched for specialized metabolism. In another recent study by Boutanaev *et al.*,[9] it has been found that, in a larger number of plant genomes, the genes encoding terpene synthases and cytochrome P450s are frequently clustered. Nonetheless, it is still largely unclear how widespread the phenomenon of metabolic gene clustering is throughout the plant kingdom, how such clusters originate, and to what extent they are maintained during evolution.

Besides co-expression and genomic clustering, there are two other important strategies to identify biosynthetic pathways: evolutionary genomic approaches that use phylogenetic profiling to look at co-occurrence across genomes or that identify recent gene family expansions, and epigenomic approaches that harness shared patterns in chromatin-level regulation based on histone modification data (Fig. 1). In this review, we will discuss each of these four strategies, with respect to both chromosomally clustered and non-clustered pathways. Subsequently, we will discuss ways in which candidate pathways can be prioritized for experimental characterization, and how synthetic biology approaches can be applied to heterologously express the pathways to identify novel natural products.

2. Plant biosynthetic pathways: clustered and non-clustered

More than two dozen plant biosynthetic pathways have now been shown to be encoded in gene clusters,[5–7] which gives useful insights into the variations in cluster architecture (Fig. 2). The plant metabolic gene clusters reported so far range in size from ~35 kb to several hundred kb and consist of three to ten genes. We define gene clusters as genomic loci that include genes for a minimum of at least three (and sometimes six or more) different types of biosynthetic reactions (*i.e.* genes encoding functionally different (sub)classes of enzymes).[5,7] These distinctions are likely to turn out to be somewhat arbitrary as we learn more about the nature of metabolic gene clusters in plants and the birth, life and death of these forms of genomic organization.

Dr Marnix H. Medema is an Assistant Professor of Bioinformatics at Wageningen University, The Netherlands. He has a BSc. in Biology and an MSc. in Biomolecular Sciences. He completed his PhD in 2013 in the Takano and Breitling groups at the University of Groningen. During his PhD he also spent time in the Fischbach lab at University of California, San Francisco. Following a postdoctoral fellowship at the Max Planck Institute for Marine Microbiology in Bremen, Germany, he joined Wageningen University in 2015, where his group develops and uses state-of-the-art computational tools to understand specialized metabolism from a genomic perspective.

Anne Osbourn is a project leader at the John Innes Centre, an honorary professor at the University of East Anglia, and Director of the Norwich Research Park Industrial Biotechnology Alliance. Her research focuses on plant-derived natural products – function, synthesis, and mechanisms of metabolic diversification. An important advance from the Osbourn laboratory has been the discovery that genes for specialized metabolic pathways are organized in 'operon-like' clusters in plant genomes, a finding that has opened up new opportunities for elucidation of new pathways and chemistries through genome mining. Anne also developed and leads SAW, a cross-curricular science education programme (http://www.sawtrust.org).

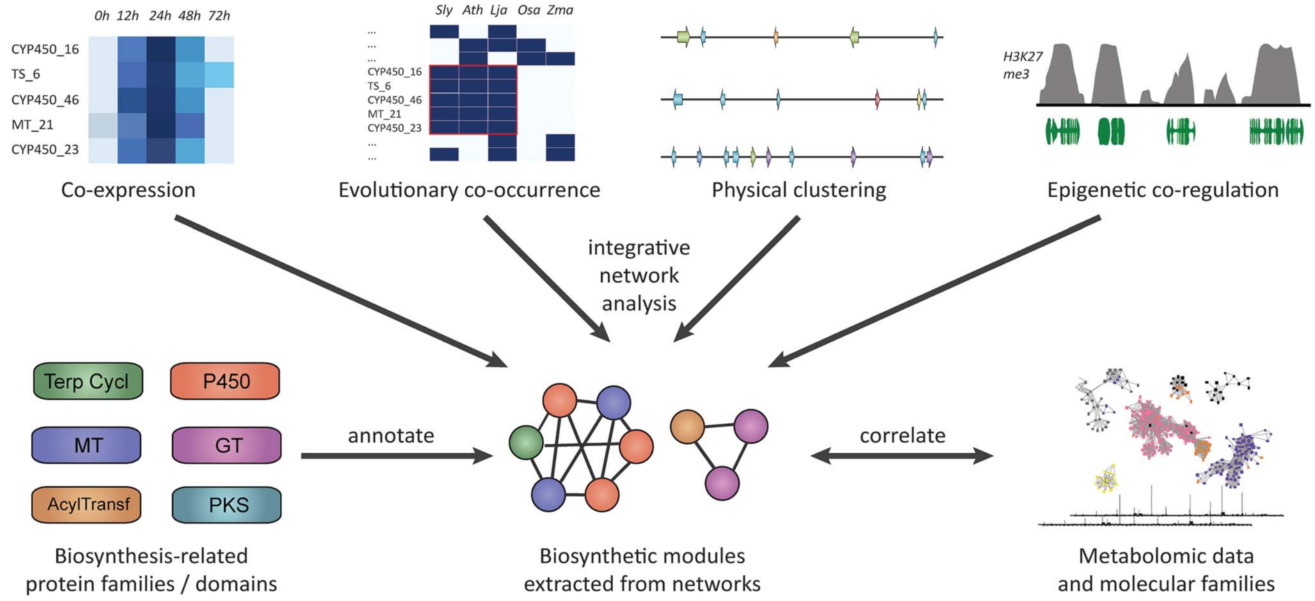


Fig. 1 Approaches for plant biosynthetic pathway discovery. Physical co-clustering, co-expression, evolutionary co-occurrence, and epigenetic co-regulation can all be used to identify candidate biosynthetic pathways. Using, for example, network analysis, these approaches can also be combined, if sufficient data is available. Functionally cohesive modules can then be extracted from such a network and annotated for the presence of genes encoding biosynthesis-related protein families. Finally, modules that have a strong biosynthetic signature can be correlated to metabolite counts or molecular families derived from molecular networking[82,83] of metabolite data.

A comprehensive picture of how widespread the phenomenon of gene clustering is in plants is still lacking. For several known molecules, the genes that encode their biosynthetic enzymes are not clustered. For example, the strigolactone biosynthesis genes *MAX1*, *MAX3*, *MAX4* and *AtD27* (ref. 10) are dispersed across three chromosomes of the *A. thaliana* genome, *MAX1* and *MAX3* being 8 Mb apart on chromosome 2, and *MAX4* and *AtD27* being located on chromosomes 4 and 1, respectively. Similarly, the biosynthetic pathways for glucosinolates,[11] arabidopyrones,[12] 4-hydroxyindole-3-carbonyl nitrile[3] and camalexin[13] in *A. thaliana* and the pathways for hydroxy-cinnamic acid amide,[14] acylsugars[15,16] and oxylipin[17] in tomato show no or very limited clustering of their biosynthetic genes. Many of these pathways are non-linear, having branches to alternative end products, which, as we will see further below, might be one potential reason why no clustering is observed.

In some other cases, there is evidence of partial clustering of different types of pathway genes. For example, the genes encoding the enzymes for the first two steps in betalain biosynthesis (a CYP76 cytochrome P450 enzyme that coverts tyrosine to L-DOPA and an L-DOPA 4,5-dioxygenase) lie within 50 kb of each other in sugar beet (*Beta vulgaris*) on chromosome 2. A betalain regulatory locus (the R locus) is also linked to the CYP76 locus.[18] The gene encoding the third step of the pathway, the UDP-glucosyltransferase cDOPA5GT, lies on chromosome 1 and is unlinked to these first two steps. This pathway does not therefore represent a clear-cut example of a clustered metabolic pathway. Nevertheless, the clustering of genes for two different steps in a metabolic pathway involving entirely different classes of enzymes is interesting. Terpene synthase and cytochrome P450 genes have also been found to be clustered in the conifer genome, suggesting that these types of association may have some functional significance.[19,20] Another good example of partial clustering is the monoterpene indole alkaloid biosynthetic pathway from *Catharanthus roseus*.[21] This pathway has been shown to involve multiple cases where two different steps are encoded on the same genomic locus; in fact, this local clustering greatly aided the elucidation of some steps in the pathway. It is not fully certain yet, however, whether or not some of these loci are part of one or more larger gene clusters, as the current genome assembly is still quite fragmented.

For those clusters that meet the definition of a plant natural product biosynthetic gene cluster (see above), some contain all of the pathway genes. Examples include the thalianol and marneral clusters in *A. thaliana*, which consist of four and three genes, respectively,[22,23] as well as the three-gene cyanogenic glycoside clusters in sorghum, *Lotus japonicus* and cassava, each of which has evolved independently.[24] In barley, a three-gene cluster has been identified that is necessary and sufficient for the biosynthesis of polyketide diketones that determine the *Cer-cqu* waxy leaf phenotype.[25] In other cases, while the core cluster contains most of the genes in the pathway, there are some anomalies. For example, the ten-gene cluster for the synthesis of the medicinal benzylisoquinoline alkaloid noscapine in poppy (*Papaver somniferum*; Fig. 3) contains all of the pathway genes except the gene for a *cis-N*-methyltransferase (TNMT).[26] Since the genome sequence of poppy is not available, this cluster was defined by assembly of a bacterial artificial chromosome (BAC) contig spanning the genes encoding the pathway enzymes. Although the location of the *TNMT* gene has not been absolutely defined, *TNMT* gene homologues are present in the flanking regions of the sequenced cluster region. In other cases some of the pathway genes are less tightly linked to the main cluster region and there may be intervening genes

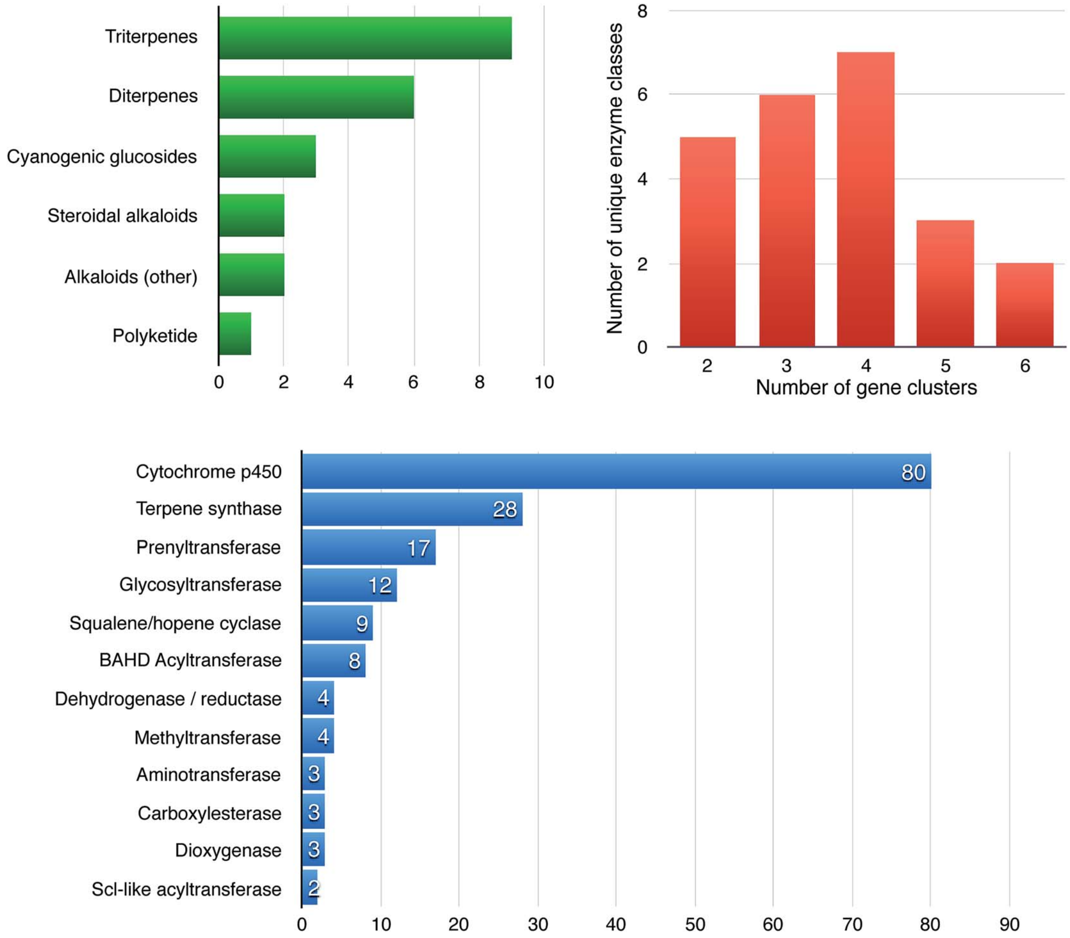


Fig. 2 Features and statistics of 28 known plant biosynthetic gene clusters. The graphs show the distributions of compound classes produced from known enzymes encoded in plant biosynthetic gene clusters (green), the number of unique (broad) enzyme families per gene cluster (red) and the gene counts of enzyme families across all clusters (blue). The numbers for the latter two are based on automated annotation of broad enzyme families through the Pfam database;[84] it should therefore be noted that any two enzymes from one Pfam protein family can still catalyze two significantly different chemical reactions. In all specific cases where only two enzyme classes are present in a cluster according to the figure, one of these comprises multiple distinct subclasses of cytochrome P450s belonging to at least two different P450 subfamilies.

with no obvious function in specialized metabolism between these 'peripheral' genes and the core cluster. For example, seven of the genes for the biosynthesis of benzoxazinoids in maize form a cluster on chromosome four[27–30] (Fig. 3). A further pathway gene encoding an *O*-methyltransferase (*Bx7*) is loosely linked, lying within 15 Mb of the core cluster. The core cluster contains a gene encoding the sugar transferase Bx8, which is required for benzoxazinoid glucosylation. A *Bx8* homologue *Bx9*, encoding an enzyme that also has activity towards benzoxazinoids, is located on a different chromosome, although it is not known whether Bx9 is a *bona fide* part of the benzoxazinoid pathway in maize. In diploid oat (*Avena strigosa*), five genes for the biosynthesis of the antimicrobial triterpene glycoside avenacin are located within a 200-kb region that does not contain any other obvious intervening genes (Fig. 3). These genes encode enzymes required for the synthesis, oxidation and acylation of the triterpene scaffold.[31–34] Two other loci (*Sad3* and *Sad4*) have been shown by mutation to be required for avenacin glycosylation but have not yet been cloned.[35] *Sad3* is loosely linked to the avenacin cluster (within 3.6 centimorgans); *Sad4* is unlinked. However, *Sad4* is required for glycosylation of other compounds in addition to avenacins and is not absolutely required for avenacin biosynthesis. Sad4 therefore appears to be 'moonlighting' and is not an integral part of the pathway. The above examples illustrate the importance of having rigorous support for the involvement of candidate pathway components in the synthesis of particular natural products in plants. *In vitro* assays of enzyme function can be misleading. Enzymatic information generated *in vitro* may therefore result in inappropriate implication of enzymes that are not dedicated pathway components.

Examples of metabolic gene clusters that contain genes encoding at least three different types of metabolic enzymes are given above. In some cases, the cognate pathways also involve one or more steps encoded by occasional 'peripheral' pathway genes, although as stated above the interpretation of which genes are *bone fide* pathway genes and which ones encode moonlighting enzymes that are not genuine pathway components requires some consideration. In other cases, there is a core cluster of genes encoding pathway enzymes but there are

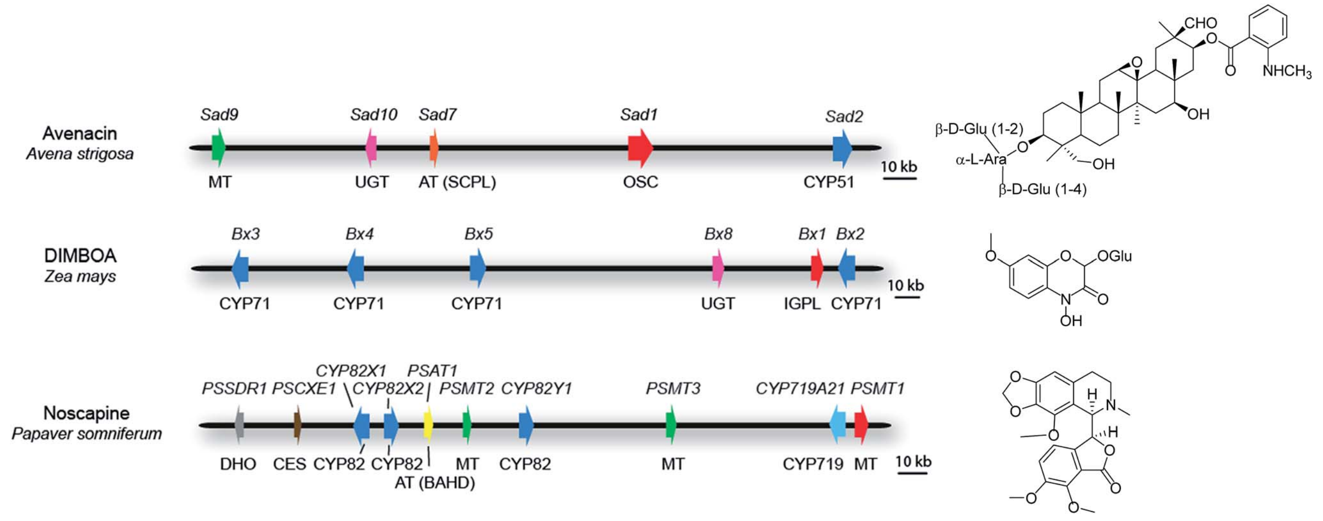

Avenacin
Avena strigosa
Sad9 *Sad10* *Sad7* *Sad1* *Sad2*
MT UGT AT (SCPL) OSC CYP51
10 kb
β-D-Glu (1-2)
α-L-Ara
β-D-Glu (1-4)
CHO
NHCH3
OH
DIMBOA
Zea mays
Bx3 *Bx4* *Bx5* *Bx8* *Bx1* *Bx2*
CYP71 CYP71 CYP71 UGT IGPL CYP71
10 kb
OGlu
Noscapine
Papaver somniferum
PSSDR1 *PSCXE1* *CYP82X1* *CYP82X2* *PSAT1* *PSMT2* *CYP82Y1* *PSMT3* *CYP719A21* *PSMT1*
DHO CES CYP82 CYP82 MT CYP82 MT CYP719 MT
AT (BAHD)
10 kb

Fig. 3 Some examples of plant metabolic gene clusters. Genes are indicated by arrows and gene(s) for the first committed pathway step are indicated in red. Gene names are indicated above the clusters and class of biosynthetic enzyme below. Abbreviations: OSC, oxidosqualene cyclase; IGPL, indole 3-glycerol phosphate lyase; AT (SCPL), SCPL-acyltransferase; AT (BAHD), BAHD-acyltransferase; MT, methyltransferase; UGT, UDP-dependent sugar transferase; DHO, dehydrogenase/reductase; CES, carboxylesterase; CYP, cytochrome P450. The oat avenacin cluster contains five genes for the synthesis, oxidation and acylation of the triterpene scaffold.[31–34] Two other loci (*Sad3* and *Sad4*) have been shown to be required for avenacin glucosylation but not yet cloned. *Sad3* lies within 3.6 cM of the core cluster while *Sad4* is unlinked.[35,85] The maize DIMBOA pathway includes three genes that are not shown in the figure; *Bx7*, which is separated from the core cluster by an intervening region of 15 Mb; the sugar transferase gene *Bx9*, which is located on a different chromosome; finally a further gene *Bx6* is not shown because its genomic location has not yet been established.[27–30] The noscapine cluster from poppy contains all of the pathway genes except the gene for tetrahydroprotoberberine *cis*-*N*-methyltransferase (TNMT), which catalyses the first committed pathway step.[26]

also additional 'mini-clusters' of up to two genes encoding steps required for elaboration of the pathway scaffold.[7] Examples of the latter include the pathways for the biosynthesis of steroidal glycoalkaloids in tomato and potato. In tomato, most of the genes for steroidal glycoalkaloid biosynthesis are clustered on chromosome seven. Two further pathway genes are adjacent to each other on a different chromosome. A similar situation exists in the close relative of tomato, potato, where regions syntenic to those in tomato contain the genes required for the biosynthesis of the steriodal alkaloids alpha-chaconine and alpha-solanine.[36] Similarly, in cucumber (*Cucumis sativus*) five genes encoding the enzymes required for the biosynthesis of triterpene glycosides associated with bitterness (cucurbitacins) are clustered on chromosome six, while four other genes are located elsewhere in the genome – a CYP71 and two CYP88 genes are co-located on chromosome three, and a CYP87 gene on chromosome one.[37]

Finally, things are also complicated by the fact that many complex plant genome assemblies are either not available or highly fragmented. For highly fragmented genome sequences, clustering will not be evident if the genes are separated on different contigs in the assembly. Where genome sequences are not available and analysis has relied on exploitation of transcriptome data, it is simply not known whether pathway genes are clustered or not. For example, the steps in the mayapple podophyllotoxin pathway[2] have been elucidated but the genome sequence has not been determined. Similarly, the artemisinin pathway has been well characterized but the genome sequence of the producing plant *Artemisia annua* is not available. As more (complex) plant genomes are sequenced, we will learn more about the diversity and variation of biosynthetic gene clustering in plants.

3. Computational identification of plant biosynthetic pathways

3.1 Identifying plant gene clusters

For the discovery of clustered pathways, much can be learned from previous work on bacterial and fungal biosynthetic pathway discovery. Various algorithms for the identification of biosynthetic gene clusters have been designed for use on bacterial and fungal genome sequences.[38–45] However, the implemented detection logic for these algorithms makes several assumptions that do not hold true for plant genomes. Specifically, most assume that (1) every time a gene encoding a 'scaffold-generating' enzyme such as a polyketide or terpene synthase is identified in a genome, it will be surrounded by a gene cluster and that (2) all clustered groups of biosynthetic genes in a genome will encode a multi-step biosynthetic pathway. On the contrary, some plant genes that encode scaffold-generating enzymes occur either as singletons (without a surrounding gene cluster) or in tandem arrays of nearly identical copies that do not encode subsequent enzymatic steps in a pathway. To further complicate things, some tandem arrays, such as the array of Bx2-3-4-5 cytochrome P450-encoding genes in the DIMBOA gene cluster[27] in fact do encode subsequent steps in a pathway that have evolved through multiple iterations of duplication and divergence.

Also, many (but not all[46]) microbe-oriented algorithms assume that all biosynthetic gene clusters will contain a gene encoding a 'scaffold-generating' enzyme (which, for example, produces a peptide, polyketide or terpene backbone). However, some plant 'scaffold-generating' reactions are in fact modifications of primary metabolic 'scaffolds' such as amino acids[47] and are therefore not necessarily linked to one specific 'scaffold-generating' enzyme family. A further complication in the process of identifying plant biosynthetic gene clusters is that in some cases the pathways are split over multiple loci, such as in the cases of tomatine/solanine[36] and cucurbitacin;[37] hence, even when a given biosynthetic pathway involves a scaffold-generating enzyme such as a terpene synthase, not all loci that code for steps of that pathway will necessarily encode one.

To overcome these challenges, algorithms for the detection of plant biosynthetic pathways should aim to identify *all* genes encoding biosynthetic enzymes (*i.e.*, as many as possible), instead of just those encoding the scaffold-generating enzymes. For the identification of plant gene clusters, this would require constructing a carefully curated and comprehensive catalogue of sequence models (*e.g.*, profile hidden Markov models[48]) for the detection of enzyme-coding genes involved in specialized metabolism. Additionally, when encountering genomic loci encoding enzymes of the same enzyme superfamily, intelligent checks should be implemented to evaluate whether these enzymes are sufficiently different from one another to be likely to catalyze different reactions, for example by evaluating their overall amino acid sequence similarity or even the similarity of amino acids surrounding the enzyme active site (based on known structural models for homologous proteins): while it is possible that two terpene synthases with 90% amino acid sequence identity and identical active site residues have functionally diverged through changes in only one or two crucial amino acids, the presence of different catalytic functions would be much more likely if they only had 30% sequence identity and major differences in their active sites. Also, detailed computational subclassification of broad enzyme superfamilies such as cytochrome P450s into their constituent families would greatly help to both identify the presence of distinct subclasses in a genomic locus and to predict their potential functions in the encoded biosynthetic pathway.[7]

Besides biological differences between plants and microbes, there are also practical differences: for example, plant genes can have very complex intron–exon structures that make it difficult to correctly predict the protein-coding regions of a genome. Most of the time, this problem is overcome by making use of RNA-Seq datasets to identify the exons; however, many biosynthetic pathways (*e.g.*, for defense-related compounds) are not expressed under typical conditions, which potentially leads to large-scale misannotations of precisely such important genomic features. Especially for non-model plants, it might therefore be important to, *e.g.*, evaluate multiple possible gene models when identifying the presence of key biosynthetic domains in a genomic region.

Regardless of the important differences between microbial and plant biosynthetic gene clusters, many of the principles found in tools like antiSMASH[40–42] can potentially be adapted to the unique properties of natural product biosynthesis in the plant kingdom. These principles include, *e.g.*, (1) the detection of enzyme families through the use of profile hidden Markov models,[48] constructed from multiple sequence alignments, that capture the sequence diversity of a protein family and can be used to recognize additional members of such a family, (2) the use of pattern matching and machine learning to assess enzyme active sites based on 3D information from crystal structures,[49] which can be used to predict activity and substrate specificity of key biosynthetic enzymes (although a meta-analysis of plant terpene synthases in 2009 found few correlations between residues and catalytic mechanisms[50]), and (3) comparative genomic analysis of gene clusters in different organisms[51,52] to infer functions of conserved parts of biosynthetic pathways from homology.

3.2 Using epigenetics and co-expression analysis to identify clustered and unclustered pathways

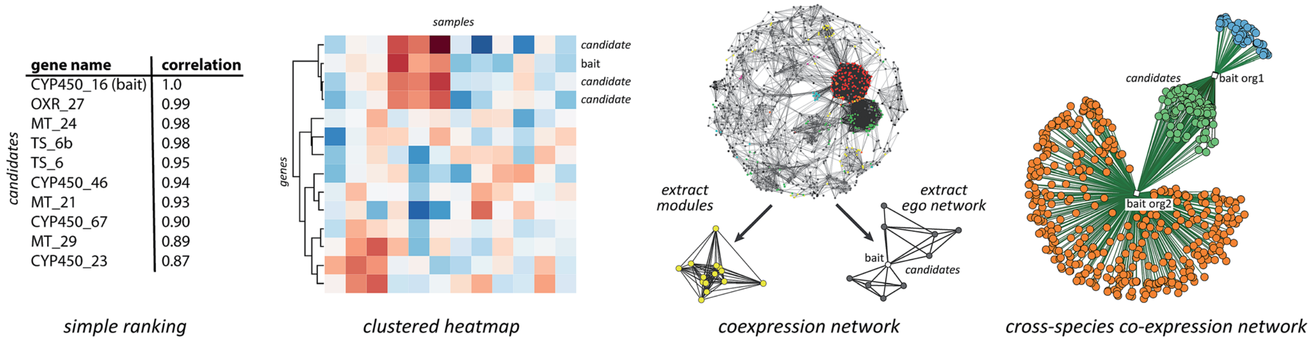
The genes for specialized metabolic pathways in plants are under strict regulatory control. Specialized metabolites are often synthesized only in particular cell types, at certain developmental stages and/or in response to environmental triggers such as pest and pathogen attack or elicitor treatment. Consistent with this, Omranian *et al.*[53] found that, in the data of Chae *et al.*,[8] genes for specialized metabolic pathways are more often coexpressed than genes encoding enzymes involved in non-specialized metabolism (having an assortativity statistic—which measures to which extent the corresponding nodes in a coexpression network are connected—of 0.118 *versus* 0.066, $P < 0.001$). Clearly, the co-ordinate transcription of genes depends on their availability to pathway-specific transcription factors. The organization of pathway genes in physically linked metabolic clusters has the potential to provide an additional higher level of regulation above and beyond that of unlinked pathways through condensation/decondensation of localized cluster-wide chromatin domains.[5–7] When the chromatin encompassing a cluster is decondensed into euchromatin, the pathway genes will become accessible for transcription. In contrast, when the cluster region is sequestered in condensed heterochromatin, it will be inaccessible, thus providing protection against ectopic expression and production of potentially toxic metabolites and pathway end products. Such a mechanism could be regarded as a 'belt and braces' approach to the regulation of these recently evolved and highly insulated clustered pathways – a way of dealing with dangerous new chemistry.

Several lines of evidence indicate that plant metabolic gene clusters are subject to regulation at the level of chromatin. DNA fluorescence *in situ* hybridization (DNA FISH) using probes for genes within the oat avenacin cluster indicates that the cluster region undergoes chromatin decondensation when in its active form.[54] Investigation of the regulation of the thalianol and marneral clusters in *A. thaliana* using chromatin mutants and chromatin immunoprecipitation (ChIP) has implicated two major chromatin markings in cluster regulation: the histone 2 variant H2A.Z in cluster activation, mediated by the SWR1 chromatin remodeling complex;[55] and polycomb-mediated histone H3 lysine 27 trimethylation (H3K27me) in cluster repression.[7] Thus, H2A.Z and H3K27me3 appear to be involved in a dynamic transition between different chromatin environments associated with the active and inactive states of the clusters. Biosynthetic gene clusters from oat, maize and rice also have pronounced H3K27me3 marking when in their 'off' state, suggesting that polycomb-mediated cluster repression occurs in both eudicots and monocots. These chromatin markings delineate the clusters and are discrete, encompassing the biosynthetic genes but not the immediate functionally unrelated flanking genes. Knowledge of these hallmark features has opened up new strategies for genome mining and has enabled the discovery of novel biosynthetic gene clusters.[56] In contrast, the genes of non-clustered plant biosynthetic pathways examined so far do not have pronounced H3K27me3 or H2A.Z markings.

It is not yet clear whether the genes within individual clustered specialized metabolic pathways are more highly co-expressed than those of non-clustered specialized metabolic pathways. In any case, co-expression studies (Fig. 4) have contributed greatly to the elucidation of non-clustered pathways. For example, the 4-hydroxyindole-3-carbonyl nitrile pathway in *A. thaliana* was recently elucidated by using a single cytochrome P450-encoding gene as 'bait' to identify other genes that were strongly co-expressed with it.[3] In a similar way, the podophyllotoxin pathway from mayapple was identified in the same lab,[2] although in this case genome sequence information is not available in order to know whether the pathway is clustered or not.

Besides co-expression analysis using bait genes or co-expression clustering, co-expression networks are also a popular tool to identify expression patterns that can lead to the discovery of biosynthetic pathways.[87] In such a network, each gene is represented by a node, and the nodes of genes whose expression shows a correlation above a certain threshold are connected by an edge. If metabolomic data is available from the same experiment, metabolites can also be added to the same network with a separate node type. The cutoff that is used to determine which connections should be shown in such a network is often arbitrary, but an approach like random matrix theory[57] can be used to rationalize this. The disadvantage of co-expression networks is that, in many experiments, large numbers of genes are correlated to each other, leading to a big 'hairball' that is difficult to interpret. This can be alleviated by, for example, first performing differential expression analysis on a biological induction or tissue where certain pathways are expected to be overexpressed, and then generating a co-expression network for only the differentially expressed genes. Alternatively, the complexity of the network can be reduced by only considering genes with protein domains related to specialized metabolism. Another option is to use the bait gene approach within the network to extract 'ego networks': a first-order ego network, for example, contains all the direct neighbours in the network of the bait gene, while a second-order ego network will also contain the direct neighbours of these neighbours. Finally, algorithms like weighted correlation network analysis (WGCNA),[86] the Markov Cluster algorithm (MCL)[58] or Infomap[59] can be used to break up a complex network into small clusters that can be individually studied for the presence of biosynthesis-related domains or correlated with metabolite absence/presence under the same conditions.

Normally, a co-expression network is constructed based on expression data from a range of different tissues and/or biological or physical treatments. An alternative to this, however, is to exploit the evolutionary variation in expression between species: during evolution, genes involved in the same pathway are expected to remain co-expressed, while the expression of other genes and pathways are expected to slowly diverge from it. When detailed orthology predictions are made for all the genes in multiple related genomes, they can be combined in a cross-species co-expression network to aid in pathway identification. The gene clusters for α-tomatine and α-solanine in tomato and potato were discovered based on such a cross-species co-expression analysis, where it was found that a range of genes co-expressed with a bait gene were clustered on two chromosomes of each of the species.[36] In plant taxa with many sequenced genomes and transcriptomes, such as the Solanaceae and Brassicaceae, such an analysis could also be potentially applied across many species at once. An online tool, CoExpNetViz,[60] has recently been published that can be used for bait-driven analysis of such cross-species networks.



gene name	correlation
CYP450_16 (bait)	1.0
OXR_27	0.99
MT_24	0.98
TS_6b	0.98
TS_6	0.95
CYP450_46	0.94
MT_21	0.93
CYP450_67	0.90
MT_29	0.89
CYP450_23	0.87

Fig. 4 Co-expression techniques to identify biosynthetic pathway components. The simplest way to identify novel candidates for a pathway is to use a bait gene that is known to be involved in the pathway and to rank all other genes by correlation coefficient to the bait. In order to also visualize the interrelationships between all (relevant) genes, clustered heatmaps can be used. The same is true for coexpression networks, which have the added advantage that they can also be used in 'untargeted' approaches to identify candidate pathways by extracting modules out of the network without using a bait. Finally, cross-species co-expression networks can be used to identify orthologous groups of genes whose co-expression is conserved over longer evolutionary periods.

3.3 Evidence from evolutionary genomics for pathway prediction

Another interesting and largely unexploited opportunity to predict functional connections between biosynthetic genes is the use of evolutionary genomic analysis. After all, genes whose products are involved in the same biochemical process are expected to co-evolve in terms of gain/loss events. For years, phylogenetic profiling techniques[61] have exploited this feature to predict biological pathways from absence/presence matrices of genes/proteins across large numbers of species. Recently, an advanced version of such an algorithm was published (termed 'CLIME'), which uses a tree-structured hidden Markov model to infer the evolutionary histories of genes and to subsequently predict evolutionarily conserved multi-gene modules based on this.[62] Given the way the algorithm works, it will probably be most effective to predict (or aid in the prediction of) linear multi-enzyme pathways that are either entirely present or entirely absent in most species. If a pathway is more branched and has many biochemical variations across species, theory predicts that this will likely blot out the absence/presence-based signal. However, even for more branched pathways, evolutionary analysis can be very useful: if it is known that certain metabolites are taxon-specific, it is expected that at least some of the corresponding biosynthetic enzymes will also be encoded by taxon-specific genes. This is exemplified by how the caffeine biosynthetic pathway was largely uncovered thanks to the discovery of a species-specific gene family expansion of *N*-methyltransferases,[63] which through duplication and divergence allowed precursors to be transformed into this well-known tri-methylated compound. Similarly, 'blooms' of cytochrome P450 diversification have been associated with the evolution of specialized metabolic diversity.[64] For the study of species for which genome and metabolite information of close relatives is available, this principle could potentially be utilized in an automated manner for multiple natural products in parallel. While the information by itself might not be sufficient to unequivocally pinpoint a gene-metabolite connection, it does have significant potential to add statistical power to clustering-based or co-expression-based approaches when included in an integrative bioinformatic framework (Fig. 1).

4. Prioritization and functional reconstitution of plant pathways

4.1 The need to prioritize

Several advances in sequencing technologies are currently making it increasingly affordable and straightforward to sequence complete plant genomes. Long read technologies like those of Pacific Biosciences and artificial long read technologies like Illumina TruSeq and 10X Genomics offer large improvements in the ability to close complex plant genomes.[65] Moreover, recently revived optical mapping technologies such as those from BioNano Genomics potentially allow scaffolding of resulting contigs to close chromosomes from end to end. Accordingly, a high-quality *de novo* assembly of a medium-sized plant genome can now be obtained for merely $30–50k, which is a fraction of past prices. As further technological improvements are still rapidly ongoing, it should soon be feasible to sequence genomes from hundreds and thousands of plant species, including a wide range of plants that have been used as herbal medicines for centuries. These will encode very large numbers of biosynthetic pathways and gene clusters, which means that computational tools will be of great importance to classify and cluster these pathways and to prioritize them for experimental characterization to drive natural product discovery. One powerful technique that has been successfully used to navigate through large numbers of biosynthetic pathways in bacteria is the use of similarity networks and pathway family reconstruction.[46,66,67] These techniques make it possible to effectively visualize biosynthetic diversity, identify novel classes of pathways and correlate pathways to metabolomic and phenotypic data. Similar tools could be developed for predicted plant pathways, by adapting the single-locus-based approaches from bacteria to accommodate both single-locus and multi-locus pathways predicted to be encoded in plant genomes. At the same time, the expected accumulation of sequenced genomes will allow unprecedented opportunities to analyze the evolution of biosynthetic pathways and gene clusters, in order to understand how the stunning molecular diversity of plant chemistry worldwide has come about over millions of years.

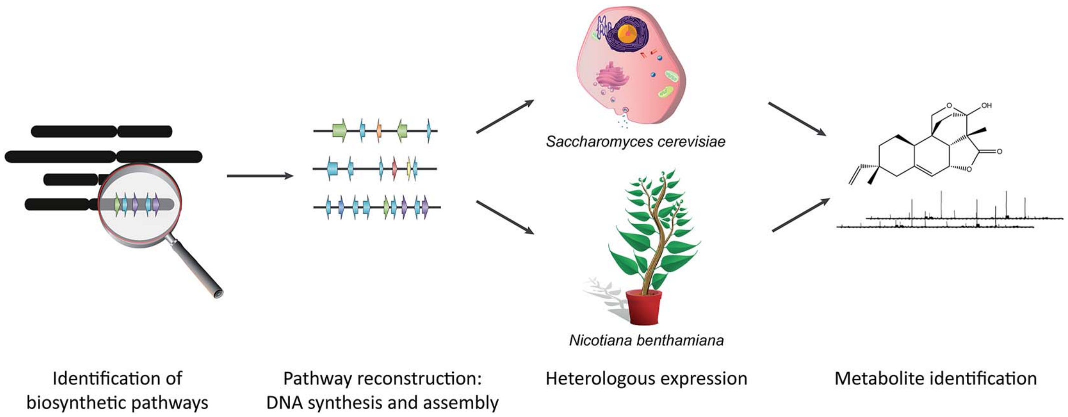

Saccharomyces cerevisiae
OH
O
O
O
Nicotiana benthamiana
Identification of biosynthetic pathways
Pathway reconstruction: DNA synthesis and assembly
Heterologous expression
Metabolite identification

Fig. 5 Synthetic biology approaches to characterize plant biosynthetic pathways. For identified (candidate) pathways, a construct is synthesized and assembled that contains all genes needed to produce the end product of the pathway, as well as the required regulatory elements. The construct is then expressed in either yeast or tobacco, after which the metabolite is identified and further characterized.

4.2 Synthetic biology technologies for functional reconstitution of candidate pathways

Rapid advances in sequencing technology, the availability of an ever-growing number of genome sequences from diverse plant species, and the application of powerful computational tools for discovery of biosynthetic gene clusters is expected to yield huge numbers of candidate pathways that will require functional validation. Synthetic biology technologies (Fig. 5) are well equipped to deal with this. The genes of new predicted biosynthetic clusters can be synthesized by commercial DNA synthesis companies and the cost of DNA synthesis is likely to decrease going forwards. DNA assembly methods that enable multiple parts to be assembled in a single reaction mean that construction of expression vectors is no longer a rate-limiting step. Assembly in yeast normally relies on overlap-dependent recombination. Golden Gate assembly methods based on Type IIS restriction enzymes have been widely adopted for plant synthetic biology and a common syntax has been proposed to enable exchange of compatible DNA parts.[68–73] Highly complex pathways for the synthesis of plant natural products have been reconstructed successfully in yeast, including those for the synthesis of artemisinic acid from wormwood, opioids from poppy and the monoterpene indole alkaloid strictosidine from Madagascar periwinkle.[74–76] Plant-based expression systems have the potential to overcome problems that may be encountered with post-translational processing, subcellular localization, precursor supply, sequestration and toxicity in yeast. Transient expression in leaves of the wild tobacco species *Nicotiana benthamiana* is proving to be a highly effective system for expression of single and multiple biosynthetic pathway genes.[77–81] This method involves infiltrating *N. benthamiana* leaves with *A. tumefaciens* containing the appropriate expression vectors and is very rapid, yielding results within around a week.

5. Conclusions

The discover-build-test cycle for plant biosynthetic pathways and gene clusters will inevitably accelerate with further technological advances and access to a rapidly growing number of plant genome sequences, including for medicinal plants. Computational genomic analysis will be particularly important in enabling the vast metabolic potential of plant genomes to be unlocked by opening up doors that lead to previously unexplored reservoirs of new enzymes, pathways and chemistries. This rapidly growing body of knowledge will feed back into and inform the continued development of computational tools for genomics-based natural product discovery in plants.

6. Acknowledgements

M. H. M. is supported by VENI grant 863.15.002 from The Netherlands Organization for Scientific Research (NWO). A. O. is supported by the UK Biotechnological and Biological Sciences Research Council (BBSRC) Institute Strategic Programme Grant ‘Understanding and Exploiting Plant and Microbial Metabolism’ (BB/J004561/1), the John Innes Foundation, the joint Engineering and Physical Sciences Research Council/BBSRC-funded OpenPlant Synthetic Biology Research Centre grant BB/L014130/1 and a National Institutes of Health Genome to Natural Products Network award U101GM110699. We thank Hernando Suarez for calculating the statistics behind Fig. 2 and Hans-Wilhelm Nützmann for preparing a draft of Fig. 5.

7. References

1 F. Geu-Flores, N. H. Sherden, V. Courdavault, V. Burlat, W. S. Glenn, C. Wu, E. Nims, Y. Cui and S. E. O'Connor, *Nature*, 2012, **492**, 138–142.

2 W. Lau and E. S. Sattely, *Science*, 2015, **349**, 1224–1228.

3 J. Rajniak, B. Barco, N. K. Clay and E. S. Sattely, *Nature*, 2015, **525**, 376–379.

4 M. H. Medema and M. A. Fischbach, *Nat. Chem. Biol.*, 2015, **11**, 639–648.

5 H.-W. Nützmann and A. Osbourn, *Curr. Opin. Biotechnol.*, 2014, **26**, 91–99.

6 S. Boycheva, L. Daviet, J.-L. Wolfender and T. B. Fitzpatrick, *Trends Plant Sci.*, 2014, **19**, 447–459.

7 H. W. Nützmann, A. Huang and A. Osbourn, *New Phytol.*, 2016, DOI: 10.1111/nph.13981.

8 L. Chae, T. Kim, R. Nilo-Poyanco and S. Y. Rhee, *Science*, 2014, **344**, 510–513.

9 A. M. Boutanaev, T. Moses, J. Zi, D. R. Nelson, S. T. Mugford, R. J. Peters and A. Osbourn, *Proc. Natl. Acad. Sci. U. S. A.*, 2014, **112**, E81–E88.

10 C. Ruyter-Spira, S. Al-Babili, S. van der Krol and H. Bouwmeester, *Trends Plant Sci.*, 2013, **18**, 72–83.

11 I. E. Sønderby, F. Geu-Flores and B. A. Halkier, *Trends Plant Sci.*, 2010, **15**, 283–290.

12 J.-K. Weng, Y. Li, H. Mo and C. Chapple, *Science*, 2012, **337**, 960–964.

13 A. P. Klein, G. Anarat-Cappillino and E. S. Sattely, *Angew. Chem., Int. Ed. Engl.*, 2013, **52**, 13625–13628.

14 E. Von Roepenack-Lahaye, M.-A. Newman, S. Schornack, K. E. Hammond-Kosack, T. Lahaye, J. D. G. Jones, M. J. Daniels and J. M. Dow, *J. Biol. Chem.*, 2003, **278**, 43373–43383.

15 A. L. Schilmiller, G. D. Moghe, P. Fan, B. Ghosh, J. Ning, A. D. Jones and R. L. Last, *Plant Cell*, 2015, **27**, 1002–1017.

16 P. Fan, A. M. Miller, A. L. Schilmiller, X. Liu, I. Ofner, A. D. Jones, D. Zamir and R. L. Last, *Proc. Natl. Acad. Sci. U. S. A.*, 2016, **113**, E239–E248.

17 G. A. Howe, G. I. Lee, A. Itoh, L. Li and A. E. DeRocher, *Plant Physiol.*, 2000, **123**, 711–724.

18 S. F. Brockington, Y. Yang, F. Gandia-Herrero, S. Covshoff, J. M. Hibberd, R. F. Sage, G. K. S. Wong, M. J. Moore and S. A. Smith, *New Phytol.*, 2015, **207**, 1170–1180.

19 C. I. Keeling and J. Bohlmann, *New Phytol.*, 2006, **170**, 657–675.

20 D. E. Hall, P. Zerbe, S. Jancsik, A. L. Quesada, H. Dullat, L. L. Madilao, M. Yuen and J. Bohlmann, *Plant Physiol.*, 2013, **161**, 600–616.

21 F. Kellner, J. Kim, B. J. Clavijo, J. P. Hamilton, K. L. Childs, B. Vaillancourt, J. Cepela, M. Habermann, B. Steuernagel,

L. Clissold, K. McLay, C. R. Buell and S. E. O'Connor, *Plant J.*, 2015, **82**, 680–692.

22 B. Field and A. E. Osbourn, *Science*, 2008, **320**, 543–547.

23 B. Field, A.-S. Fiston-Lavier, A. Kemen, K. Geisler, H. Quesneville and A. E. Osbourn, *Proc. Natl. Acad. Sci. U. S. A.*, 2011, **108**, 16116–16121.

24 A. M. Takos, C. Knudsen, D. Lai, R. Kannangara, L. Mikkelsen, M. S. Motawia, C. E. Olsen, S. Sato, S. Tabata, K. Jørgensen, B. L. Møller and F. Rook, *Plant J.*, 2011, **68**, 273–286.

25 L. M. Schneider, N. M. Adamski, C. E. Christensen, D. B. Stuart, S. Vautrin, M. Hansson, C. Uauy and P. von Wettstein-Knowles, *J. Exp. Bot.*, 2016, **67**, 2715–2730.

26 T. Winzer, V. Gazda, Z. He, F. Kaminski, M. Kern, T. R. Larson, Y. Li, F. Meade, R. Teodor, F. E. Vaistij, C. Walker, T. A. Bowser and I. A. Graham, *Science*, 2012, **336**, 1704–1708.

27 M. Frey, P. Chomet, E. Glawischnig, C. Stettner, S. Grün, A. Winklmair, W. Eisenreich, A. Bacher, R. B. Meeley, S. P. Briggs, K. Simcox and A. Gierl, *Science*, 1997, **277**, 696–699.

28 U. von Rad, R. Hüttl, F. Lottspeich, A. Gierl and M. Frey, *Plant J.*, 2001, **28**, 633–642.

29 M. Frey, K. Huber, W. J. Park, D. Sicker, P. Lindberg, R. B. Meeley, C. R. Simmons, N. Yalpani and A. Gierl, *Phytochemistry*, 2003, **62**, 371–376.

30 R. Jonczyk, H. Schmidt, A. Osterrieder, A. Fiesselmann, K. Schullehner, M. Haslbeck, D. Sicker, D. Hofmann, N. Yalpani, C. Simmons, M. Frey and A. Gierl, *Plant Physiol.*, 2008, **146**, 1053–1063.

31 X. Qi, S. Bakht, M. Leggett, C. Maxwell, R. Melton and A. Osbourn, *Proc. Natl. Acad. Sci. U. S. A.*, 2004, **101**, 8233–8238.

32 X. Qi, S. Bakht, B. Qin, M. Leggett, A. Hemmings, F. Mellon, J. Eagles, D. Werck-Reichhart, H. Schaller, A. Lesot, R. Melton and A. Osbourn, *Proc. Natl. Acad. Sci. U. S. A.*, 2006, **103**, 18848–18853.

33 S. T. Mugford, X. Qi, S. Bakht, L. Hill, E. Wegel, R. K. Hughes, K. Papadopoulou, R. Melton, M. Philo, F. Sainsbury, G. P. Lomonossoff, A. D. Roy, R. J. M. Goss and A. Osbourn, *Plant Cell*, 2009, **21**, 2473–2484.

34 S. T. Mugford, T. Louveau, R. Melton, X. Qi, S. Bakht, L. Hill, T. Tsurushima, S. Honkanen, S. J. Rosser, G. P. Lomonossoff and A. Osbourn, *Plant Cell*, 2013, **25**, 1078–1092.

35 P. Mylona, A. Owatworakit, K. Papadopoulou, H. Jenner, B. Qin, K. Findlay, L. Hill, X. Qi, S. Bakht, R. Melton and A. Osbourn, *Plant Cell*, 2008, **20**, 201–212.

36 M. Itkin, U. Heinig, O. Tzfadia, A. J. Bhide, B. Shinde, P. D. Cardenas, S. E. Bocobza, T. Unger, S. Malitsky, R. Finkers, Y. Tikunov, A. Bovy, Y. Chikate, P. Singh, I. Rogachev, J. Beekwilder, A. P. Giri and A. Aharoni, *Science*, 2013, **341**, 175–179.

37 Y. Shang, Y. Ma, Y. Zhou, H. Zhang, L. Duan, H. Chen, J. Zeng, Q. Zhou, S. Wang, W. Gu, M. Liu, J. Ren, X. Gu, S. Zhang, Y. Wang, K. Yasukawa, H. J. Bouwmeester, X. Qi, Z. Zhang, W. J. Lucas and S. Huang, *Science*, 2014, **346**, 1084–1088.

38 M. H. Li, P. M. Ung, J. Zajkowski, S. Garneau-Tsodikova and D. H. Sherman, *BMC Bioinf.*, 2009, **10**, 185.

39 N. Khaldi, F. T. Seifuddin, G. Turner, D. Haft, W. C. Nierman, K. H. Wolfe and N. D. Fedorova, *Fungal Genet. Biol.*, 2010, **47**, 736–741.

40 M. H. Medema, K. Blin, P. Cimermancic, V. de Jager, P. Zakrzewski, M. A. Fischbach, T. Weber, E. Takano and R. Breitling, *Nucleic Acids Res.*, 2011, **39**, W339–W346.

41 K. Blin, M. H. Medema, D. Kazempour, M. A. Fischbach, R. Breitling, E. Takano and T. Weber, *Nucleic Acids Res.*, 2013, **41**, W204–W212.

42 T. Weber, K. Blin, S. Duddela, D. Krug, H. U. Kim, R. Bruccoleri, S. Y. Lee, M. A. Fischbach, R. Müller, W. Wohlleben, R. Breitling, E. Takano and M. H. Medema, *Nucleic Acids Res.*, 2015, **43**, W237–W243.

43 M. A. Skinnider, C. A. Dejong, P. N. Rees, C. W. Johnston, H. Li, A. L. H. Webster, M. A. Wyatt and N. A. Magarvey, *Nucleic Acids Res.*, 2015, **43**, 9645–9662.

44 C. W. Johnston, M. A. Skinnider, M. A. Wyatt, X. Li, M. R. M. Ranieri, L. Yang, D. L. Zechel, B. Ma and N. A. Magarvey, *Nat. Commun.*, 2015, **6**, 8421.

45 T. Wolf, V. Shelest, N. Nath and E. Shelest, *Bioinformatics*, 2015, **32**, 1138–1143.

46 P. Cimermancic, M. H. Medema, J. Claesen, K. Kurita, L. C. Wieland Brown, K. Mavrommatis, A. Pati, P. A. Godfrey, M. Koehrsen, J. Clardy, B. W. Birren, E. Takano, A. Sali, R. G. Linington and M. A. Fischbach, *Cell*, 2014, **158**, 412–421.

47 G. Anarat-Cappillino and E. S. Sattely, *Curr. Opin. Plant Biol.*, 2014, **19**, 51–58.

48 S. R. Eddy, *Bioinformatics*, 1998, **14**, 755–763.

49 M. Röttig, C. Rausch and O. Kohlbacher, *PLoS Comput. Biol.*, 2010, **6**, e1000636.

50 J. Degenhardt, T. G. Köllner and J. Gershenzon, *Phytochemistry*, 2009, **70**, 1621–1637.

51 M. H. Medema, E. Takano and R. Breitling, *Mol. Biol. Evol.*, 2013, **30**, 1218–1223.

52 M. Hadjithomas, I.-M. A. Chen, K. Chu, A. Ratner, K. Palaniappan, E. Szeto, J. Huang, T. B. K. Reddy, P. Cimermančič, M. A. Fischbach, N. N. Ivanova, V. M. Markowitz, N. C. Kyrpides and A. Pati, *mBio*, 2015, **6**, e00932.

53 N. Omranian, S. Kleessen, T. Tohge, S. Klie, G. Basler, B. Mueller-Roeber, A. R. Fernie and Z. Nikoloski, *Trends Plant Sci.*, 2015, **20**, 266–268.

54 E. Wegel, R. Koumproglou, P. Shaw and A. Osbourn, *Plant Cell*, 2009, **21**, 3926–3936.

55 H.-W. Nützmann and A. Osbourn, *New Phytol.*, 2015, **205**, 503–510.

56 N. Yu, H.-W. Nützmann, J. T. MacDonald, B. Moore, B. Field, S. Berriri, M. Trick, S. J. Rosser, S. V. Kumar, P. S. Freemont and A. Osbourn, *Nucleic Acids Res.*, 2016, **44**, 2255–2265.

57 F. Luo, Y. Yang, J. Zhong, H. Gao, L. Khan, D. K. Thompson and J. Zhou, *BMC Bioinf.*, 2007, **8**, 299.

58 A. J. Enright, S. Van Dongen and C. A. Ouzounis, *Nucleic Acids Res.*, 2002, **30**, 1575–1584.

59 M. Rosvall and C. T. Bergstrom, *Proc. Natl. Acad. Sci. U. S. A.*, 2008, **105**, 1118–1123.

60 O. Tzfadia, T. Diels, S. De Meyer, K. Vandepoele, A. Aharoni and Y. Van de Peer, *Front. Plant Sci.*, 2016, **6**, 1194.

61 M. Pellegrini, E. M. Marcotte, M. J. Thompson, D. Eisenberg and T. O. Yeates, *Proc. Natl. Acad. Sci. U. S. A.*, 1999, **96**, 4285–4288.

62 Y. Li, S. E. Calvo, R. Gutman, J. S. Liu and V. K. Mootha, *Cell*, 2014, **158**, 213–225.

63 F. Denoeud, L. Carretero-Paulet, A. Dereeper, G. Droc, R. Guyot, M. Pietrella, C. Zheng, A. Alberti, F. Anthony, G. Aprea, J.-M. Aury, P. Bento, M. Bernard, S. Bocs, C. Campa, A. Cenci, M.-C. Combes, D. Crouzillat, C. Da Silva, L. Daddiego, F. De Bellis, S. Dussert, O. Garsmeur, T. Gayraud, V. Guignon, K. Jahn, V. Jamilloux, T. Joët, K. Labadie, T. Lan, J. Leclercq, M. Lepelley, T. Leroy, L.-T. Li, P. Librado, L. Lopez, A. Muñoz, B. Noel, A. Pallavicini, G. Perrotta, V. Poncet, D. Pot, Priyono, M. Rigoreau, M. Rouard, J. Rozas, C. Tranchant-Dubreuil, R. VanBuren, Q. Zhang, A. C. Andrade, X. Argout, B. Bertrand, A. de Kochko, G. Graziosi, R. J. Henry, Jayarama, R. Ming, C. Nagai, S. Rounsley, D. Sankoff, G. Giuliano, V. A. Albert, P. Wincker and P. Lashermes, *Science*, 2014, **345**, 1181–1184.

64 B. Hamberger and S. Bak, *Philos. Trans. R. Soc. London, Ser. B*, 2013, **368**, 20120426.

65 R. VanBuren, D. Bryant, P. P. Edger, H. Tang, D. Burgess, D. Challabathula, K. Spittle, R. Hall, J. Gu, E. Lyons, M. Freeling, D. Bartels, B. Ten Hallers, A. Hastie, T. P. Michael and T. C. Mockler, *Nature*, 2015, **527**, 508–511.

66 N. Ziemert, A. Lechner, M. Wietz, N. Millán-Aguiñaga, K. L. Chavarria and P. R. Jensen, *Proc. Natl. Acad. Sci. U. S. A.*, 2014, **111**, E1130–E1139.

67 J. R. Doroghazi, J. C. Albright, A. W. Goering, K.-S. Ju, R. R. Haines, K. A. Tchalukov, D. P. Labeda, N. L. Kelleher and W. W. Metcalf, *Nat. Chem. Biol.*, 2014, **10**, 963–968.

68 T. Ellis, T. Adie and G. S. Baldwin, *Integr. Biol.*, 2011, **3**, 109–118.

69 W. Liu, J. S. Yuan and C. N. Stewart, *Nat. Rev. Genet.*, 2013, **14**, 781–793.

70 N. J. Patron, *Curr. Opin. Plant Biol.*, 2014, **19**, 14–19.

71 N. J. Patron, D. Orzaez, S. Marillonnet, H. Warzecha, C. Matthewman, M. Youles, O. Raitskin, A. Leveau, G. Farré, C. Rogers, A. Smith, J. Hibberd, A. A. R. Webb, J. Locke, S. Schornack, J. Ajioka, D. C. Baulcombe, C. Zipfel, S. Kamoun, J. D. G. Jones, H. Kuhn, S. Robatzek, H. P. Van Esse, D. Sanders, G. Oldroyd, C. Martin, R. Field, S. O'Connor, S. Fox, B. Wulff, B. Miller, A. Breakspear, G. Radhakrishnan, P.-M. Delaux, D. Loqué, A. Granell, A. Tissier, P. Shih, T. P. Brutnell, W. P. Quick, H. Rischer, P. D. Fraser, A. Aharoni, C. Raines, P. F. South, J.-M. Ané, B. R. Hamberger, J. Langdale, J. Stougaard, H. Bouwmeester, M. Udvardi, J. A. H. Murray, V. Ntoukakis, P. Schäfer, K. Denby, K. J. Edwards, A. Osbourn and J. Haseloff, *New Phytol.*, 2015, **208**, 13–19.

72 A. Casini, M. Storch, G. S. Baldwin and T. Ellis, *Nat. Rev. Mol. Cell Biol.*, 2015, **16**, 568–576.

73 M. H. Medema, R. Kottmann, P. Yilmaz, M. Cummings, J. B. Biggins, K. Blin, I. de Bruijn, Y. H. Chooi, J. Claesen, R. C. Coates, P. Cruz-Morales, S. Duddela, S. Düsterhus, D. J. Edwards, D. P. Fewer, N. Garg, C. Geiger, J. P. Gomez-Escribano, A. Greule, M. Hadjithomas, A. S. Haines, E. J. N. Helfrich, M. L. Hillwig, K. Ishida, A. C. Jones, C. S. Jones, K. Jungmann, C. Kegler, H. U. Kim, P. Kötter, D. Krug, J. Masschelein, A. V. Melnik, S. M. Mantovani, E. A. Monroe, M. Moore, N. Moss, H.-W. Nützmann, G. Pan, A. Pati, D. Petras, F. J. Reen, F. Rosconi, Z. Rui, Z. Tian, N. J. Tobias, Y. Tsunematsu, P. Wiemann, E. Wyckoff, X. Yan, G. Yim, F. Yu, Y. Xie, B. Aigle, A. K. Apel, C. J. Balibar, E. P. Balskus, F. Barona-Gómez, A. Bechthold, H. B. Bode, R. Borriss, S. F. Brady, A. A. Brakhage, P. Caffrey, Y.-Q. Cheng, J. Clardy, R. J. Cox, R. De Mot, S. Donadio, M. S. Donia, W. A. van der Donk, P. C. Dorrestein, S. Doyle, A. J. M. Driessen, M. Ehling-Schulz, K.-D. Entian, M. A. Fischbach, L. Gerwick, W. H. Gerwick, H. Gross, B. Gust, C. Hertweck, M. Höfte, S. E. Jensen, J. Ju, L. Katz, L. Kaysser, J. L. Klassen, N. P. Keller, J. Kormanec, O. P. Kuipers, T. Kuzuyama, N. C. Kyrpides, H.-J. Kwon, S. Lautru, R. Lavigne, C. Y. Lee, B. Linquan, X. Liu, W. Liu, A. Luzhetskyy, T. Mahmud, Y. Mast, C. Méndez, M. Metsä-Ketelä, J. Micklefield, D. A. Mitchell, B. S. Moore, L. M. Moreira, R. Müller, B. A. Neilan, M. Nett, J. Nielsen, F. O'Gara, H. Oikawa, A. Osbourn, M. S. Osburne, B. Ostash, S. M. Payne, J.-L. Pernodet, M. Petricek, J. Piel, O. Ploux, J. M. Raaijmakers, J. A. Salas, E. K. Schmitt, B. Scott, R. F. Seipke, B. Shen, D. H. Sherman, K. Sivonen, M. J. Smanski, M. Sosio, E. Stegmann, R. D. Süssmuth, K. Tahlan, C. M. Thomas, Y. Tang, A. W. Truman, M. Viaud, J. D. Walton, C. T. Walsh, T. Weber, G. P. van Wezel, B. Wilkinson, J. M. Willey, W. Wohlleben, G. D. Wright, N. Ziemert, C. Zhang, S. B. Zotchev, R. Breitling, E. Takano and F. O. Glöckner, *Nat. Chem. Biol.*, 2015, **11**, 625–631.

74 S. Brown, M. Clastre, V. Courdavault and S. E. O'Connor, *Proc. Natl. Acad. Sci. U. S. A.*, 2015, **112**, 3205–3210.

75 S. Galanie, K. Thodey, I. J. Trenchard, M. Filsinger Interrante and C. D. Smolke, *Science*, 2015, **349**, 1095–1100.

76 C. J. Paddon, P. J. Westfall, D. J. Pitera, K. Benjamin, K. Fisher, D. McPhee, M. D. Leavell, A. Tai, A. Main, D. Eng, D. R. Polichuk, K. H. Teoh, D. W. Reed, T. Treynor, J. Lenihan, M. Fleck, S. Bajad, G. Dang, D. Dengrove, D. Diola, G. Dorin, K. W. Ellens, S. Fickes, J. Galazzo, S. P. Gaucher, T. Geistlinger, R. Henry, M. Hepp, T. Horning, T. Iqbal, H. Jiang, L. Kizer, B. Lieu, D. Melis, N. Moss, R. Regentin, S. Secrest, H. Tsuruta, R. Vazquez, L. F. Westblade, L. Xu, M. Yu, Y. Zhang, L. Zhao, J. Lievense, P. S. Covello, J. D. Keasling, K. K. Reiling, N. S. Renninger and J. D. Newman, *Nature*, 2013, **496**, 528–532.

77 K. Geisler, R. K. Hughes, F. Sainsbury, G. P. Lomonossoff, M. Rejzek, S. Fairhurst, C.-E. Olsen, M. S. Motawia, R. E. Melton, A. M. Hemmings, S. Bak and A. Osbourn, *Proc. Natl. Acad. Sci. U. S. A.*, 2013, **110**, E3360–E3367.

78 K. Miettinen, L. Dong, N. Navrot, T. Schneider, V. Burlat, J. Pollier, L. Woittiez, S. van der Krol, R. Lugan, T. Ilc, R. Verpoorte, K.-M. Oksman-Caldentey, E. Martinoia, H. Bouwmeester, A. Goossens, J. Memelink and D. Werck-Reichhart, *Nat. Commun.*, 2014, **5**, 3606.

79 R. Thimmappa, K. Geisler, T. Louveau, P. O'Maille and A. Osbourn, *Annu. Rev. Plant Biol.*, 2014, **65**, 225–257.

80 P. Saxena, E. C. Thuenemann, F. Sainsbury and G. P. Lomonossoff, *Methods Mol. Biol.*, 2016, **1385**, 39–54.

81 C. Crocoll, N. Mirza, M. Reichelt, J. Gershenzon and B. A. Halkier, *Frontiers in Bioengineering and Biotechnology*, 2016, **4**, 14.

82 J. Watrous, P. Roach, T. Alexandrov, B. S. Heath, J. Y. Yang, R. D. Kersten, M. van der Voort, K. Pogliano, H. Gross, J. M. Raaijmakers, B. S. Moore, J. Laskin, N. Bandeira and P. C. Dorrestein, *Proc. Natl. Acad. Sci. U. S. A.*, 2012, **109**, E1743–E1752.

83 D. D. Nguyen, C. H. Wu, W. J. Moree, A. Lamsa, M. H. Medema, X. Zhao, R. G. Gavilan, M. Aparicio, L. Atencio, C. Jackson, J. Ballesteros, J. Sanchez, J. D. Watrous, V. V. Phelan, C. van de Wiel, R. D. Kersten, S. Mehnaz, R. de Mot, E. A. Shank, P. Charusanti, H. Nagarajan, B. M. Duggan, B. S. Moore, N. Bandeira, B. Ø. Palsson, K. Pogliano, M. Gutierrez and P. C. Dorrestein, *Proc. Natl. Acad. Sci. U. S. A.*, 2013, **110**, E2611–E2620.

84 R. D. Finn, A. Bateman, J. Clements, P. Coggill, R. Y. Eberhardt, S. R. Eddy, A. Heger, K. Hetherington, L. Holm, J. Mistry, E. L. L. Sonnhammer, J. Tate and M. Punta, *Nucleic Acids Res.*, 2014, **42**, D222–D230.

85 K. Papadopoulou, R. E. Melton, M. Leggett, M. J. Daniels and A. E. Osbourn, *Proc. Natl. Acad. Sci. U. S. A.*, 1999, **96**, 12923–12928.

86 P. Langfelder and S. Horvath, *BMC Bioinformatics*, 2008, **9**, 559.

87 E. A. Serin, H. Nijveen, H. W. Hilhorst and W. Ligterink, *Front. Plant Sci.*, 2016, **7**, 444.